Lab 3: Correlation and regression I

Statistics for Social Scientists II

Bur, GJM

2024-09-23

1 Understanding correlation: working "by hand" in Stata

1. Pull the NHANES data from my Github. Drop respondents who are missing on either variable. For respondents below age 18, find the correlation between height (`bmxht`) and weight (`bmxwt`). Interpret the results.

   Time-permitting, first, do this by hand; if unable, skip to the end of this question and follow the instructions there. This is what next week's homework will cover, so you'll need to learn it either way.

   First, make a "product variable" called `htwt` and find its mean. Then substract from that the product of the individual variables' means; for large n, this is almost exactly equal to the covariance of the two variables.

   Then, find the correlation by dividing this covaraince by the product of the variables' individual standard deviations. Again, do this by "hand": successively `sum` the two variables, then make a new variable that is equal to the square of the old variable less its mean. The *mean* of this new variable will approximately be the variance of each variable, respectively. Here's some example code. Remember that lines involving locals must be run all at once to work.

```
drop if missing(<var1>, <var2>)
* make product var
gen prod = <var1>*<var2>
sum prod
local prodmean = r(mean)
sum <var1>
local mean1 = r(mean)
sum <var2>
local mean2 = r(mean)
di `prodmean' - `mean1'*`mean2'
corr <var1> <var2>, cov // checking work; if good, store in local
local cov = `prodmean' - `mean1'*`mean2'

* make variance var
```

```
sum <var1>
gen <var1>sq = (<var1> - r(mean))^2
sum <var1>sq
di sqrt(r(mean)) // this should be ~= to the sd; if it is, store in a local
local sd1 = sqrt(r(mean))

...

local r = `cov'/(`sd1'*`sd2')
di `r'
```

Now, how could you do this more quickly? First, get the covariance and variances with …

```
corr <var1> <var2>, cov
```

Then, access the elements of the matrix of results, `r(C)`. Just append the element you want with a bracket and then the location. `di r(C)[2,1] / sqrt(r(C)[1,1] * r(C)[2,2])` Even quicker, as you might suspect, is just `corr <var1> <var2>`.

2 Understanding regression: working "by hand" in Stata

2.1 Regression from $\hat{\sigma}_{X,Y}$ and $\hat{\sigma}_X$

2. Find the regression slope and intercept for the regression of weight on height. *Do this using the covariance and the variance of X*. I'll show the standard syntax later. Interpret the line.

2.2 Standardizing, regression slopes, standardizing regression slopes,

3. Standardize each variable, making note as you go of the original $\hat{\sigma}$ of each variable.

   OK, we probably get the point with hand calculation. Good to do once in a while to keep your brain in shape. Let's do this more easily. *See if you can find the syntax for standardizing this from my posted do-file from last week (I put it up over the weekend).* You can do it with a loop or not.

4. Using the results you found previously for the slope (check your work with `reg <yvar> <xvar>`), find the regression slope.

   Big hint, and an answer to those of you who asked "when will we ever use the summation properties you showed us": right now, and many times later. Recall that the sample covariance is…

$$\hat{\sigma}_{X,Y} = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})(y_j - \bar{y})$$

So, therefore...

$$\begin{aligned}
\hat{\sigma}_{aX,bY} &= \frac{1}{n-1}\sum_{j=1}^{n}(ax_j - a\bar{x})(by_j - b\bar{y}) \\
&= \frac{1}{n-1}\sum_{j=1}^{n} ab(x_j - \bar{x})(y_j - \bar{y}) \\
&= ab \cdot \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})(y_j - \bar{y}) \\
&= ab \cdot \hat{\sigma}_{X,Y}
\end{aligned}$$

So the numerator of the bivariate regression slope is multiplied by $\frac{1}{\hat{\sigma}_x \hat{\sigma}_y}$. What happens to the denominator?

$$\begin{aligned}
\hat{\sigma}_X^2 &= \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})^2 \\
\hat{\sigma}_{aX}^2 &= \frac{1}{n-1}\sum_{j=1}^{n}(ax_j - a\bar{x})^2 \\
&= \frac{1}{n-1}\sum_{j=1}^{n} a^2(x_j - \bar{x})^2 \\
&= a^2 \cdot \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})^2 \\
&= a^2 \cdot \hat{\sigma}_X^2
\end{aligned}$$

So the denominator of the bivariate regression slope is multiplied by $\frac{1}{\hat{\sigma}_X^2}$, so that term ends up in the numerator. How does the whole thing change, then?

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\hat{\sigma}_{X,Y}}{\hat{\sigma}_X^2} \\
\hat{\beta}_{1,std.} &= \frac{\hat{\sigma}_X^2}{\hat{\sigma}_X \hat{\sigma}_Y}\frac{\hat{\sigma}_{X,Y}}{\hat{\sigma}_X^2} \\
\hat{\beta}_{1,std.} &= \frac{1}{\hat{\sigma}_X \hat{\sigma}_Y}\hat{\sigma}_{X,Y} \\
&= r
\end{aligned}$$

5. Convert height to inches and weight to pounds. What do you predict will happen to the *original* slope? Use the properties given above. By the way, what would happen if

we looked at the correlation/standardized slope—any change? Why or why not? You can work out the answer to this from the properties of variance and covariance given above, as well as the fact that $r = \frac{\hat{\sigma}_{X,Y}}{\hat{\sigma}_X \hat{\sigma}_Y}$.

2.3 Standard error for regression slopes

6. Time-permitting: find the standard error for the regression slope by hand, and then use it to construct a 95 CI and a two-tailed hypothesis test against $H_0 : \beta_1 = 0$.

2.4 Standard error *of* regression, AKA conditional standard deviation AKA root mean squared error

7. Time-permitting: find the conditional standard deviation of the regression. There are many ways to obtain this quantity. It is standard in Stata's output (see if you can find it), but more useful to try to calculate once. Perhaps even more instructive is make a variable for the residuals, then examine its standard deviation. The average of the quantity below should be very close to the (in-sample) mean squared error.

```
corr x y, cov
local cov = r(C)[2, 1]
local vx = r(C)[1,1]
local vy = r(C)[2,2]
local b1 = `cov'/`vx'
di `b1'
sum y
local ybar = r(mean)
sum x
local xbar = r(mean)
local b0 = `ybar' - `b1'*`xbar'

gen yhat = `b1'*x + `b0'
gen resid = (y - yhat)
sum resid // this is approximately correct, but our df is n-2
```

Faster would be

```
reg bmxwt bmxht
predict ehat, residuals
sum ehat
```